PRAVDEPODOBNOSŤ A ŠTATISTIKA

Zákony veľkých čísel

doc. RNDr. Štefan Peško, CSc.

Katedra matematických metód a operačnej analýzy, FRI ŽU

13. júna 2018

Motivácia

Ak opakujeme nezávisle určitý pokus, môžeme z pozorovaných hodnôt zostaviť rozdelenie relatívnych početností niektorej náhodnej udalosti a pokúsiť sa zistiť jej rozdelenie alebo charakteristiky.

Očakávame, že pri dodržaní istých podmienok sa s rastúcim počtom opakovaní bude empirické rozdelenie početností približovať k teoretickému rozdeleniu.

Túto myšlienku presnejšie upravujú zákony veľkých čísel.

Konvergencia podľa pravdepodobnosti

Hovoríme, že postupnosť náhodných premenných $\{X_n\}_{n=1}^{\infty}$ konverguje podľa pravdepodobnosti ku konštante c, ak pre každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathcal{P}(|X_n - c| < \varepsilon) = 1. \tag{1}$$

resp.

$$\lim_{n\to\infty} \mathcal{P}(|X_n - c| \geq \varepsilon) = 0. \tag{2}$$

Pre konvergenciu podľa pravdepodobnosti budeme používať symbol $X_n \xrightarrow{\mathcal{P}} c$.

Tvrdenie 16 (Bernoulli)

Majme postupnosť $\{X_n\}_{n=1}^{\infty}$ navzájom nezávislých náhodných premenných, ktoré majú to isté binomické rozdelenie $X_i \sim Bi(n, p)$. Potom pre každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathcal{P}\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) = 0. \tag{3}$$

Dôkaz: Pre každé n platí

$$\mathbb{E}\left(\frac{X_n}{n}\right) = \frac{np}{n} = p,$$
$$\mathbb{D}\left(\frac{X_n}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

Z Čebyšovej nerovnosti dostaneme

$$\mathcal{P}\left(\left|\frac{X_n}{n} - p\right| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2},$$

odkiaľ je tvrdenie zrejmé.

Tvrdenie 16 hovorí, že s rastúcim počtom nezávislých pokusov, postupnosť relatívnych početností výskytu náhodnej udalosti A „nejako“ konverguje k $\mathcal{P}(A)$, ak je rovnaká vo všetkých pokusoch. Zapíšme tvrdenie (3) pomocou doplnkovej udalosti, dostaneme

$$\lim_{n\to\infty} \mathcal{P}\left(\left|\frac{X_n}{n} - p\right| < \varepsilon\right) = 1.$$

Ak si zvolíme ľubovoľne malé $\varepsilon > 0$, vždy máme jednotkovú pp., že v limite sa relatívna početnosť líši od odhadovanej pp. o menej než je toto malé ε.

Tvrdenie 17 (Čebyšev)

Majme takú postupnosť $\{X_i\}_{i=1}^n$,že $\mathbb{E}(X_i) < \infty$ a $\mathbb{D}(X_i) = c < \infty$. Potom pre každé $\varepsilon > 0$ platí

$$\lim_{n\to\infty} \mathcal{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i)\right| < \varepsilon\right) = 1. \qquad (4)$$

Dôkaz:
Vzhľadom na Čebyševovu nerovnosť [1] , môžeme uvedenú pp. odhadnúť

$$\mathcal{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i)\right| < \varepsilon\right) = 1 - \frac{\sum_{i=1}^{n} \mathbb{D}(X_i)}{n^2\varepsilon^2} \geq 1 - \frac{c}{n\varepsilon^2}.$$

Po limitnom prechode $n \to \infty$ dostaneme tvrdenie.

[1] $\mathcal{P}(|Y - \mathbb{E}(Y))| \geq \beta) \leq \frac{\mathbb{D}(Y)}{\beta^2}, \forall\beta > 0, \exists E(Y), \exists D(Y)$

Pri splnení uvedených podmienok

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathcal{P}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}(X_i).$$

Ak majú naviac náhodné premenné X_i to isté rozdelenie so strednou hodnotou $\mathbb{E}(X_i) = \mu$ a konečným rozptylom $\mathbb{D}(X_i) = \sigma^2$, potom $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\mathcal{P}} \mu$.

Príklad 6.1

Vypočítajte pravdepodobnosť, že aritmetický priemer z 10 000 nezávislých meraní udáva skutočnú hodnotu meranej veličiny μ s presnosťou 0.01 ak rozptyl jednotlivých meraní nepresiahne 0.02 .

Výsledky $n = 10000$ meraní reprezentujme premennými $X_i; i = 1, \ldots, n$ s $\mathbb{E}(X_i) = \mu, c = \mathbb{D}(X_i) = 0.02$.

$$\mathcal{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| < \varepsilon\right) \geq 1 - \frac{c}{n\varepsilon^2} = 1 - \frac{0.02}{10000(0.01)^2} = 0.98.$$

Centrálna limitná veta

Podstatou všetkých formulácií centrálnej limitnej vety je tvrdenie, že súčet veľkého počtu nezávislých náhodných premenných s konečnou strednou hodnotou a konečným rozptylom má asymptoticky normálne rozdelenie.

Tento výsledok sa využíva v praxi. Ak aj skúmané náhodné veličiny v jednotlivých pokusoch nemajú normálne rozdelenie, tak s ich súčtom sa pri dostatočne veľkom počte pracuje ako s normálne rozdelenou náhodnou premennou.

Tvrdenie 18 (Linderberg-Lévy)

Majme postupnosť $\{X_i\}_{i=1}^n$ navzájom nezávislých náhodných premenných, ktoré majú rovnaký pravdepodobnostné rozdelenie, konečnú strednú hodnotu $\mathbb{E}(X_i) = \mu$ a konečný rozptyl $\mathbb{D}(X_i) = \sigma^2$. Nech $Y = \sum_{i=1}^{n} X_i$. Potom normovaná náhodná premenná

$$U = \frac{Y - \mathbb{E}(Y)}{\sqrt{\mathbb{D}(Y)}}$$

má asymptoticky štandardizované normálne rozdelenie pp. $N(0,1)$ tj. platí

$$\lim_{n\to\infty} \mathcal{P}(U < u) = \Phi(u),\ u \in \Re.$$

Tvrdenie 19 (Moivrova–Laplaceova)

Majme postupnosť $\{X_i\}_{i=1}^n$ navzájom nezávislých náhodných premenných s rovnakým rozdelením $X_i \sim Bi(n,p), 0 < p < 1$. Potom pre ľubovoľné $x \in \Re$ platí

$$\lim_{n\to\infty} \mathcal{P}\left(\frac{\sum_{i=1}^n X_i - np}{\sqrt{np(1-p)}} < x\right) = \Phi(x).$$

Poznámka:

Tvrdenie umožňuje pri veľkom počte opakovaní n aproximovať binomické rozdelenie náhodnej premennej X normálnym tj.

$$\mathcal{P}(a \le X \le b) \approx \Phi\left(\frac{b-np}{\sqrt{p(1-p)}}\right) - \Phi\left(\frac{a-np}{\sqrt{p(1-p)}}\right).$$

Príklad 6.2

Majme náhodnú premennú $Y \sim Bi(n, p)$, ktorú môžeme vyjadriť v tvare súčtu $Y = \sum_{i=1}^{n} X_i, X_i \sim A(p)$. Potom je súčet S_n

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(\frac{X_i - p}{\sqrt{p(1-p)}} \right) = \frac{Y - np}{\sqrt{np(1-p)}},$$

normovanou náhodnou premennou k binomickému rozdeleniu. Pre veľké n je $S_n \sim N(0, 1)$, čo môžeme zapísať tiež

$$Y \sim N(np, np(1-p)). \tag{5}$$

Poznámka:

Aproximácia binomického rozdelenia normálnum sa považuje za vyhovujúcu ak platí: $np(1-p) > 9$.

Príklad 6.3

Zaujíma nás neznámy podiel p osôb s krvnou skupinou A v danej populácii. U koľkých osôb musíme zistiť, či má alebo nemá skupinu A, aby sme s pp. aspoň 0.9 odhadli neznámu pp. s chybou nanajvýš 0.05?

Vyberieme z danej populácie náhodne n osôb. Náhodný počet osôb s krvnou skupinou A je náhodnou premennou $Y \sim Bi(n, p)$. Neznámy podiel p budeme odhadovať pomocou Y/n. Počet oslovených osôb zvolíme tak, aby platilo

$$\mathcal{P}\left(\left|\frac{1}{n}Y - p\right| < 0.05\right) \approx 0.9.$$

Použijeme (5) tj. $Y \sim Bi(np, np(1-p))$ a postupnými úpravami dostaneme

$$0.9 \approx \mathcal{P}\left(\left|\frac{1}{n}Y - p\right| < 0.05\right) = \mathcal{P}(-0.05n + np < Y < 0.05n + np)$$
$$= \Phi\left(\frac{0.05n}{\sqrt{np(1-p)}}\right) - \Phi\left(\frac{-0.05n}{\sqrt{np(1-p)}}\right) = 2\Phi\left(\frac{0.05n}{\sqrt{np(1-p)}}\right) - 1.$$

Teda by malo byť

$$\Phi\left(\frac{0.05n}{\sqrt{np(1-p)}}\right) \approx \frac{1+0.9}{2} = 0.95,$$

čo zaručíme voľbou

$$\frac{0.05n}{\sqrt{np(1-p)}} = \Phi^{-1}(0.95) = 1.644854.$$

Pri voľbe $p = \frac{1}{2}$ máme $p(1-p) \to max$ a tak po úprave dostaneme

$$n > 100 \cdot 1.644854^2 \approx 270.$$

Príklad 6.4

Pomocou centrálnej limitnej vety vyjadrite $\mathcal{P}(\sum_{i=1}^{n} X_i < a)$ ak sú $X_1, X_2, \ldots, X_n$ nezávislé náhodné premenné s rozdelením $N(1,4)$, resp. $A(\frac{1}{5}), R(0,2), R(-2,2)$.

Podľa tvrdenia 18 možeme odhadnúť

$$U = \sum_{i=1}^{n} X_i \sim N(n\mathbb{E}(X_1), n\mathbb{D}(X_1)) \implies \mathcal{P}(U < a) = \Phi\Big(\frac{a - n\mathbb{E}(X_1)}{\sqrt{n\mathbb{D}(X_1)}}\Big)$$

- $X_i \sim N(1,4)$: $\mathcal{P}(U < a) = \Phi\big(\frac{a-n}{2\sqrt{n}}\big)$,
- $X_i \sim A(\frac{1}{5})$: $\mathcal{P}(U < a) \approx \Phi\big(\frac{5a-n}{2\sqrt{n}}\big)$,
- $X_i \sim R(0,2)$: $\mathcal{P}(U < a) \approx \Phi\big(\frac{(a-n)\sqrt{3}}{\sqrt{n}}\big)$,
- $X_i \sim R(-2,2)$: $\mathcal{P}(U < a) \approx \Phi\big(\frac{a\sqrt{3}}{2\sqrt{n}}\big)$.

Príklad 6.5

V určitej oblasti je 3% chorých na maláriu. Aká je pravdepodobnosť, že pri kontrole 5 000 ľudí nájdeme 3 ± 0.5 chorých na maláriu.

Podiel chorých

$$\frac{S_{5000}}{5000} \sim Bi(5000, 0.03) \approx N(0.03, \frac{1}{5000} 0.03 \cdot 0.97).$$

Potom

$$\mathcal{P}\left(2.5 \leq \frac{S_{5000}}{5000} \leq 3.5\right) \approx \Phi\left(\frac{3.5 - 0.3}{\sqrt{0.00000582}}\right) - \Phi\left(\frac{2.5 - 0.3}{\sqrt{0.00000582}}\right) = 0.962$$

Cvičenie

6.1 (3b) Počet chýb na jednej strane textu má strední hodnotu 8 a rozptyl 4. Aká je pravdepodobnosť, že na 100 stranách bude menej než 750 chyb?

6.2 (3b) Odhadnite pp. s akou bude počet šestiek, ktoré padnú v 1000 nezávislých hodoch spravodlivou kockou, v intervale $\langle 147, 186 \rangle$.

6.3 (3b) V 10 000 nezávislých hodoch mincou padlo spolu 5 087× hlava. Môžeme považovať mincu za spravodlivú?

6.4 (4b) Pravdepodobnosť, že bude výrobok reklamovaný je 0.05. Aká je pp., že z 300 predaných výrobkov ich bude nanajvýš 20 reklamovaných?

6.5 (5b) Pri zostavovaní štatistického výkazu bolo spočítaných 10^3 čísel, pričom každé bolo zaokrúhlené na m desatiných miest. Chyba vznikajúca zaokrúhlením má rovnomerné rozdelenie v rozsahu $(-0.5 \cdot 10^{-m}, 0.5 \cdot 10^{-m})$. Vypočítajte hranicu, ktorú s pp. 0.99 neprekročí absolútna hodnota celkovej chyby súčtu.